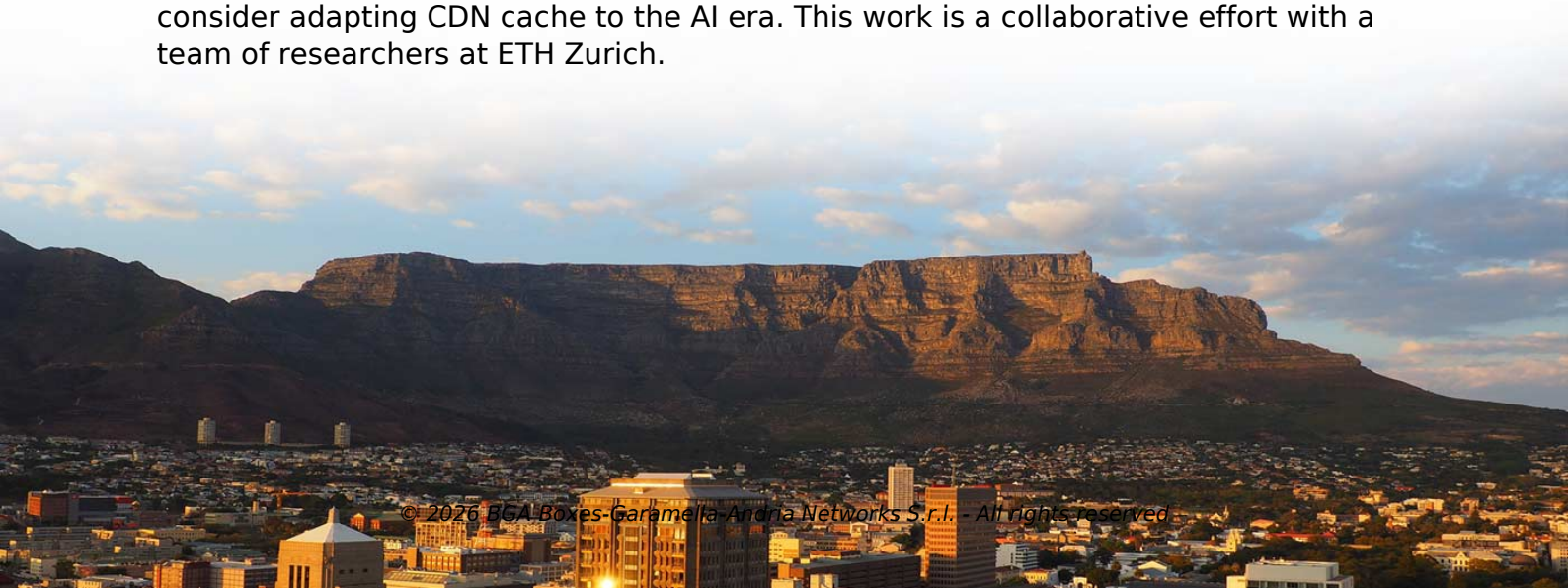
AI Router Cache Server

Overview

This article will guide you through the process of building a powerful AI proxy using NeuroLink in TypeScript, covering key components like routing, caching, rate limiting, cost tracking, and logging. Before diving into the "how," let's understand the "why. "GitHub - decolua/9router: Unlimited FREE AI coding. Connect Claude Code, Codex, Cursor, Cline, Copilot, Antigravity to FREE Claude/GPT/Gemini via 40+ providers. Auto-fallback, RTK -40% tokens, never hit limits. · GitHub feat (docker): add Docker setup, environment examples, and architectur. Update. Quick Answer: SmarterRouter (v2. 1, Feb 2026) is a Python-based LLM gateway that sits between your apps and your local backends (Ollama, llama. cpp, or OpenAI-compatible APIs). It profiles each model's speed and VRAM footprint on your hardware, then auto-routes each prompt to the best model for the. Today Red Hat announced project llm-d, a new distributed inference platform built to support our vision for an open source AI future. We have. To save on inference costs, you can enable prompt caching on supported providers and models. When using caching (whether automatically in. Emails, phone numbers, SSNs, credit cards, all scrubbed in real-time. Set per-team budgets, model. In this post, we'll explore how AI traffic impacts storage cache, describe some challenges associated with mitigating this impact, and propose directions for the community to consider adapting CDN cache to the AI era. This work is a collaborative effort with a team of researchers at ETH Zurich.



Article Content

How NVIDIA Dynamo 1.0 Powers Multi-Node Inference

NVIDIA Dynamo 1.0 delivers a mature, production-grade distributed inference framework for large-scale, multi-node AI deployments, with proven

Tom''s Hardware: For The Hardcore PC Enthusiast

The right Wi-Fi router can make a huge difference in your day-to-day productivity and gaming experience. We''ve tested a slew of models to help you

Prompt caching | OpenAI API

OpenAI routes API requests to servers that recently processed the same prompt, making it cheaper and faster than processing a prompt from scratch. Prompt Caching can reduce latency by up to 80% and

Next.js 15 | Next.js

Next.js 15 introduces React 19 support, caching improvements, a stable release for Turbopack in development, new APIs, and more.

The Best React-Based Framework | Gatsby

Gatsby is a React-based open source framework with performance, scalability and security built-in. Collaborate, build and deploy 1000x faster on Netlify.

Next.js 16 | Next.js

Next.js 16 includes Cache Components, stable Turbopack, file system caching, React Compiler support, smarter routing, new caching APIs, and React 19.2 features.

Gartner Business Insights, Strategies & Trends For

Business and Technology Insights and Trends AI''s Influence Runs Deeper Than You Think — 2026 Gartner Strategic Predictions Explain Why Understand them

Prompt caching with Azure OpenAI in Microsoft Foundry Models ...

Prompt caching reduces overall request latency and cost for longer prompts that have identical content at the beginning of the prompt. In this context, "prompt" refers to the input you send

LLM Semantic Router: Intelligent request routing for large language ...

LLM Semantic Router uses semantic understanding and caching to boost performance, cut costs, and enable efficient inference with llm-d.

Intelligent Routing for Cache Optimization | Open-Source LLM ...

This paper presents a scalable, open-source inferencing solution using the vLLM Production Stack on Dell AI Factory. It evaluates routing, observability, security, and autoscaling, demonstrating an end-to

Reduce Claude Code Tokens: 10 Tested Tools | ComputingForGeeks

Reduce Claude Code tokens 20 to 43% with 10 tested tools: token-savior, caveman, MCP caching, Haiku routing. Real before and after numbers.

Building Your Own AI Proxy: Route, Cache, and Monitor LLM

This article will guide you through the process of building a powerful AI proxy using NeuroLink in TypeScript, covering key components like routing, caching, rate limiting, cost tracking,

llm-cache-router · PyPI

Project description llm-cache-router A lightweight, production-ready Python library that combines semantic caching, multi-provider LLM routing, and cost tracking in a single async-first API. Cut your

Home :: GFI

Utilizing new AI Navigation technology, we''re redefining traditional navigation paradigms by offering a more modern approach to accessing information

Home | Buffalo Americas

I set the whole thing up on my network and then shipped the drive to the client. They plugged it into their router and installed the software and it worked like a champ. I love it when technology just works ????

Next.js 16 App Router — The Data-Fetching Mental Model

Server components, cache boundaries, dynamic vs static — the full mental model in one post.

Microsoft

Service Level Agreements (SLA) for Online Services. The Service Level Agreements (SLA) describe Microsoft''s commitments for uptime and connectivity for Microsoft Online Services

Prompt Caching

To save on inference costs, you can enable prompt caching on supported providers and models. Most providers automatically enable prompt caching, but note that some (see Alibaba and Anthropic

Unified LLM Gateway | Govern & Optimize AI Models

An AI gateway between your app and 400+ LLM providers. Change your base URL to router.requesty.ai and instantly get intelligent routing, fallbacks, cost optimization, caching, governance, and observability.

Master KV cache aware routing with llm-d for efficient

Learn how llm-d''s KV cache aware routing reduces latency and improves throughput by directing requests to pods that already hold relevant

WORLD WIDE WEB JOURNAL Home

O''Reilly & Associates, Inc. 103A Morris St. Sebastopol, CA United States

SmarterRouter: A VRAM-Aware LLM Gateway for Your Local AI Lab

Intelligent router that profiles your models, manages VRAM, caches responses semantically, and auto-picks the best model per prompt. Works with Ollama and llama.cpp.

OneUptime | The Open-Source Observability Platform

OneUptime is an open-source complete observability platform. Monitor websites, APIs, and servers. Get alerts, manage incidents, and keep customers informed

Why we''re rethinking cache for the AI era

In this post, we''ll explore how AI traffic impacts storage cache, describe some challenges associated with mitigating this impact, and propose directions for the community to consider adapting

LLM Semantic Router: Intelligent request routing for

LLM Semantic Router uses semantic understanding and caching to boost performance, cut costs, and enable efficient inference with llm-d.

Contact Us

For more information, pricing, or custom solutions, please contact us:

Website: https://www.boxesgaramella-andria.it
Email: sales@boxesgaramella-andria.it
Phone: +39 331 584 7291
Address: Via delle Industrie, 15, 20154 Milano, Italy

This document is for informational purposes only. Specifications subject to change without notice.